Behavioral Sciences and the Law
Behav. Sci. Law **28**: 106–119 (2010)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/bsl.918

Inter-Rater Reliability of the PCL-R Total and Factor Scores among Psychopathic Sex Offenders: Are Personality Features More Prone to Disagreement than Behavioral Features?

John F. Edens, Ph.D.*, Marcus T. Boccaccini, Ph.D.† and Darryl W. Johnson, Ph.D.†

Despite considerable support for the inter-rater reliability of the Hare Psychopathy Checklist—Revised (PCL-R; Hare, 1991, 2003) in research contexts, there is increasing concern that scores from this instrument may be considerably less stable across examiners in applied contexts, particularly when scoring is based on separate interviews. The present study examines archival data from a sample of imprisoned sex offenders (*n* = 20) who obtained relatively high PCL-R total scores (≥25) and were administered this instrument on a second occasion by a different examiner. Intraclass correlations for the total and Factor 2 score were lower than those generally reported in research studies. Of greater concern, Factor 1 scores were only negligibly related to each other ($ICC_{A,1} = .16$). Correcting for potential range restriction among these high scoring individuals resulted in total and Factor 2 score measures of agreement that were somewhat more consistent with published research, but Factor 1 continued to display exceedingly poor agreement across examiners. Copyright © 2010 John Wiley & Sons, Ltd.

The Hare Psychopathy Checklist—Revised (PCL-R; Hare, 1991, 2003) is a clinical rating scale that is widely used in forensic and correctional settings (Archer, Buffington-Vollum, Stredney, & Handel, 2006; DeMatteo & Edens, 2006; Lally, 2003). The academic literature on the PCL-R has recently been dominated primarily by various debates concerning the scale's factor structure (e.g., two- versus three- versus four-factor models; Cooke, Michie, & Skeem, 2007; cf. Hare & Neumann, 2005), exactly what the instrument measures (e.g., a latent taxon versus a dimensional construct; Edens, Marcus, Lilienfeld, & Poythress, 2006; cf. Harris, Rice, & Quinsey, 1994), whether, as a "personality" scale, it is excessively focused on criminal history variables (Skeem & Cooke, in press), and its validity and utility with certain groups (e.g., ethnic minorities, women, youths; Edens, Campbell, & Weir, 2007; Forth, Kosson, & Hare,

*Correspondence to: John F. Edens, Department of Psychology, Texas A&M University, 4235 TAMU, College Station, TX 77843-4235, USA. E-mail: jedens@psych.tamu.edu
†Sam Houston State University.

This research was supported by the Texas Department of Criminal Justice under a research agreement with the first author. Points of view are those of the authors and do not necessarily represent the position of the Texas Department of Criminal Justice.

Contract/grant sponsor: Texas Department of Criminal Justice.

2003; Nicholls & Petrila, 2005; Skeem, Edens, Camp, & Colwell, 2004; Sullivan & Kosson, 2006; Verona & Vitale, 2006) and/or for certain purposes (e.g., estimating risk for future violence or treatment response in various contexts; Edens, Petrila, & Buffington-Vollum, 2001; Gendreau, Goggin, & Smith, 2002; Hemphill & Hare, 2004; Skeem, Mulvey, & Monahan, 2002). Collectively, these debates broadly focus on the construct validity of scores derived from this instrument in relation to various important issues and applications (Messick, 1995)—all of which have (more or less) direct implications for forensic examiners who use this scale.

Although the validity of scores from any type of test or rating scale is certainly a critical topic to consider, validity is predicated on these scores being sufficiently reliable to allow for meaningful assessments of the variable of interest. In regards to this central psychometric question, it is common to see global references to the PCL-R as a "reliable" measure, perhaps in large part due to the accumulated evidence indicating that scores tend to be relatively internally consistent and stable across raters in published research studies (Hare, 2003; Rufino, Heinonen, Boccaccini, Murrie, & Edens, 2009). For example, the most recent edition of the PCL-R manual (Hare, 2003) reports total score intra-class correlation coefficient (ICC_1) values of .86 for 488 North American male offenders and .88 for 106 North American male forensic psychiatric patients. Additionally, the total score ICC_1 for a sample of 288 British offenders was .86. Given these values, it is perhaps not especially surprising that the reliability of PCL-R scores seems to be something of a foregone conclusion to many examiners in applied settings (at least in our experience).[1]

It should be noted that the method for establishing inter-rater reliability for this scale depends on the information available to the raters. The above-noted values for the PCL-R manual were based on reviews of institutional files and interviews with the offenders. The preferred method of scoring the items, as noted in the manual, is for the evaluator to conduct both a semi-structured interview and file review. A recent review of rater agreement findings from more than 120 PCL-R studies found that about half provided little or no information concerning whether independent interviews were used for the cases used to check rater agreement (Rufino et al., 2009). In only a handful of studies was it clear that rater agreement statistics were based on cases in which both evaluators conducted an independent interview (see, e.g., Levenson, 2004; Murrie, Boccaccini, Johnson, & Janke, 2008; Rutherford, Cacciola, Alterman, McKay, & Cook, 1999; Tyrer et al., 2005). In several of these studies, ICC values were notably lower than those reported in the PCL-R manual.

Despite the findings of generally high inter-rater reliability for the total score in most research, there have been various concerns (see, e.g., Edens, 2006; Edens & Petrila, 2006; Edens et al., 2001; Hare, 2003; Hare, 2006) expressed about the potential for PCL-R ratings in real-world cases to be influenced by several forms of bias (e.g., inadequate training, partisanship), particularly those more inferential "personality" (i.e., PCL-R Factor 1) items, such as callousness, shallow affect, superficial charm, and failure to accept responsibility for one's actions, whose rating is more heavily reliant on clinical inference and subjective judgment. Of note, there have been anecdotal accounts of exceedingly large disparities in PCL-R scores reported in some criminal cases (see,

[1] In what is perhaps an extreme example of this, DeMatteo and Edens (2006) recently described one case in which a mental health expert in a capital murder case testified that 90 out of 100 examiners would have given the defendant the exact same PCL-R score (33) that he had given him.

DOI: 10.1002/bsl

e.g., Edens & Vincent, 2008, who identified one U.S. case in which two testifying experts differed in their rating of a male prisoner by 15 points: 25 (62nd percentile) versus 10 (ninth percentile)). Although these case examples do not address which particular aspects of the PCL-R are primarily responsible for such scoring discrepancies when they occur (i.e., Factor 1, Factor 2, or some combination thereof), they do raise concerns that in at least some instances examiners largely fail to converge to any meaningful degree on the global presence/absence of psychopathic traits for a given examinee.[2] Of course, anecdotal evidence can be criticized on numerous grounds, particularly in terms of questionable representativeness of what typically occurs in most cases.

Recently, a small but growing body of research has begun to address what essentially might be referred to as the "field reliability" (Wood, Nezworski, & Stejskal, 1996) of PCL-R scores. For example, Levenson (2004) examined the stability of PCL-R scores in a sample of 69 offenders undergoing sexually violent predator (SVP) evaluations conducted in Florida who were evaluated on two separate occasions by examiners retained by a private agency contracting with the state to provide these services. Levenson reported a multiple rater ICC_1 of .84 for the PCL-R total score, which converts to a single rater value of .72 (see Boccaccini, Turner, & Murrie, 2008, p. 267), which is notably lower than the total score ICC values reported in the PCL-R manual for male prisoners (.86 to .88).

More recently, Murrie et al. (2008) compared rates of agreement between PCL-R scores provided by experts called by opposing sides (termed "petitioners" and "respondents") in Texas Sexually Violent Predator (SVP) civil commitment cases. In 23 cases, an expert retained by the respondent also administered the PCL-R. For the total score, the single-evaluator ICC_1 for absolute agreement across these cases was .39, well below the high levels of agreement observed for the PCL-R in most research contexts. Of particular note, the authors obtained mean state expert ratings of approximately 26 (SD = 8.48) and mean respondent (i.e., offender) expert ratings of approximately 18 (SD = 6.62).[3] Given that the range of PCL-R scores is only between 0 and 40 and that approximately 98% of male criminal offenders actually score between 5 and 36 (Hare, 2003, p. 164), an *average* difference of almost eight points is of considerable practical significance. Recently, Murrie et al. (2009) expanded their initial sample of 23 cases to 35 (as well as examining the reliability of other risk instruments used in SVP cases) and reported virtually identical results (ICC_1 for absolute agreement = .42). Agreement for PCL-R total scores was also low ($ICC_{A,1}$ = .47) for 22 offenders who were evaluated twice by evaluators working for the same side (i.e., petitioner; Boccaccini et al., 2008).

Results suggesting adversarial allegiance or bias in PCL-R scores are not entirely unprecedented, or limited to sexually violent predator cases. Lloyd, Clark, and Forth (in press) reported similar, although somewhat smaller, differences using scores obtained from case law reports of Canadian criminal trials. Scores from Crown experts (M = 27.32) were significantly higher than those from defense experts (M = 24.47) in

[2] Echoing these concerns, Hare (2006) recently noted "Although the research evidence for the reliability and validity of the PCL-R and its derivatives is extensive, this does not ensure that an individual assessment will be reliable or valid. In a research context, misuse of these instruments will have few negative consequences for the individual. However, when the scores are used in clinical and criminal justice contexts, the implications of misuse are potentially serious, especially if the scores are used to guide treatment or adjudication decisions."

[3] Somewhat more concretely, the petitioner and respondent mean values roughly correspond to the 67th and 32nd percentiles for U.S. male prisoners, respectively.

the 15 trials in which experts from both sides reported scores ($d = 0.58$). As a result, rater agreement for these cases was lower ($ICC_{A,1} = .67$) than the values reported in the PCL-R manual, although not as low as the values reported by Murrie et al. (2008, 2009).

In sum, there seems to be growing reason to question the stability of PCL-R scores across raters outside of academic research contexts, although considerably more research is needed on this important topic before any definitive conclusions can be drawn. One particular limitation of the available studies is that, due to the nature of the archival records available, they have been restricted to analyses of total scores and have been unable to examine what aspects of the PCL-R are more or less stable across examiners. Clearly, certain items on the PCL-R require more subjective judgment and clinical intuition (e.g., "pathological" lying, failure to accept responsibility) than do others (e.g., juvenile delinquency, criminal versatility, many short-term marital relationships). Generally speaking, it would seem that the "personality" items comprising Factor 1 of this rating scale have a greater potential for random or systematic error in their scoring (at the level of the individual examiner) than do those more behaviorally based items comprising Factor 2 (for individual item ICCs supporting this contention, see Hare, 2003, chapter 5, as well as Rutherford et al., 1999). It is noteworthy that, even in studies in which it appears that well trained examiners rely on the same file and interview data, ICCs for Factor 1 of the PCL-R tend to be lower than for Factor 2. For example, for the North American prisoner and forensic patient samples and the British prisoner samples reviewed in the PCL-R manual, the ICCs for Factor 2 were .85, .88, and .90, respectively, whereas the values for Factor 1 were .75, .79, and .74, respectively.

Some authors (Edens et al., 2001; Edens, Colwell, Desforges, & Fernandez, 2005) have argued that it is precisely the "personality" items contained within the PCL-R that have the greatest potential for abuse in real-world settings, due to the ostensibly more prejudicial nature of the traits and characteristics being quantified. Support for this argument comes both from experimental studies demonstrating the stigmatizing effects of Factor 1 traits using mock jury designs (Edens et al., 2005) and from content analyses of what occurs in real-world criminal cases. For example, in a content analysis of 20 prosecutor summations in capital sentencing hearings, Costanzo and Peterson (1994) noted that in most cases the defendant was characterized as "a cold, remorseless killer" (p. 125) and as "a liar and a manipulator" (p. 137). Jurors appear to weigh heavily such factors in their deliberations on whether to support a death sentence. In fact, California jurors participating in the Capital Jury project (Sundby, 1998) who were interviewed after imposing a death sentence used a variety of adjectives to describe defendants (e.g., cocky, emotionless, clever, lack of feeling and compassion, calculating) that could have been pulled directly from classic descriptions of the core features of the prototypical psychopath.

Given this background, the purpose of the present study is to add to the small but growing body of literature addressing the field reliability of the PCL-R by examining archival prison records to compare independent scores collected for a select sample of incarcerated sex offenders. These inmates were originally evaluated as part of a larger internal prison project intended to inform decision-making about civil commitment procedures for impending SVP legislation[4] in the state of Texas (see Edens, Hart,

[4] This legislation, which was ultimately passed into law and includes direct reference to a requirement of the assessment of psychopathy in these evaluations, is described in detail in the Texas Health & Safety Code (2000).

DOI: 10.1002/bsl

Johnson, Johnson, & Olver, 2000, for a complete description of this sample). Those individuals within this larger sample who obtained elevated scores on the PCL-R (defined as a raw score $\geq 25^5$) were subsequently referred for a second evaluation conducted by an independent examiner. As such, the re-assessments conducted on this subgroup of high scoring individuals ($n = 20$) provide an opportunity to consider the inter-reliability of PCL-R scores among a subgroup of "psychopathic" offenders—a group of particular interest to researchers, clinicians, and legal decision-makers. Importantly, this study also moves beyond earlier field reliability research by examining rater agreement for the factor scores of the PCL-R. The availability of these scores allows us to examine whether the performance of the total score may be more or less a function of ostensibly more subjective and inferential Factor 1 item content.

METHOD

Archival Offender Data

In 1998, the Texas Department of Criminal Justice initiated a study evaluating 300 incarcerated sex offenders to inform pending legislation regarding criteria to be used to identify inmates eligible for possible civil commitment as "sexually violent predators." These individuals were given full disclosure regarding the nature of the research and were informed that they could refuse participation at any point without repercussions. Of the original 300 cases, 58 were identified as possibly "high risk" for re-offending and were referred for a more extensive evaluation that included the PCL-R. As noted above, 20 of these individuals scored ≥ 25 and were subsequently re-assessed a second time by another examiner. In terms of demographics, most were white ($n = 12$, 60%), with others identified as black ($n = 6$, 30%), Hispanic ($n = 1$. 2.5%), or other ($n = 1$, 2.5%). Mean age for the participants was 29.20 (SD = 9.92). Regarding criminal history, 13 (65%) were identified as offending against children and 7 (35%) as offending against adults. Average number of prior offenses ranged from 0 to 6 ($M = 2.30$, SD = 2.23).

The PCL-R was administered according to the procedures outlined in the test manual by trained doctoral-level clinicians working at state-operated psychiatric hospitals. Ratings were based on a review of institutional files, National Crime Information Center (NCIC) reports, Texas Crime Information Center (TCIC) reports, and clinical interviews. Examiners who performed the re-assessments relied on the same file data provided to the initial examiners but an independent clinical interview was conducted for each re-assessment. Of note, the second examiners were aware that the individuals being re-evaluated had previously obtained a PCL-R score of 25 or higher.[6] The average length of time between the first and second PCL-R evaluations was 7.55 days (SD = 8.61; range = 0–32).

[5] Historically, scores ≥ 30 have been used to categorize individuals as "psychopaths" on the PCL-R (Hare, 2003). However, this is a relatively arbitrary cut point (Edens et al., 2006) and various research teams (e.g., Rutherford et al., 1999) have employed somewhat lower values (e.g., 25) to operationalize "high scores" on this rating scale.

[6] Although this obviously raises concerns about inflated reliability estimates due to some degree of knowledge regarding the outcome of the first evaluation, it seems likely that this may mimic some real-world cases in which a second evaluator (e.g., an expert retained by the defense/respondent) is aware of a score previously provided by an expert (e.g., one employed or retained by the state/petitioner).

DOI: 10.1002/bsl

Procedure

To obtain the archival data reported in this study, on-site prison files were reviewed by the research team to collect the PCL-R scores and other information (e.g., demographics and criminal history), which was entered into a laptop computer. Of note, although item-level data were available for most of the original 58 initial assessments described by Edens et al. (2000), file information available for the second assessment only reported PCL-R total score and factor scores.[7] Approval to access these archival records was provided both by the Texas prison system as well as the university institutional review board of the first author.

RESULTS

Mean Differences from First to Second Evaluation

Descriptive statistics for both assessments are reported in Table 1.[8] Perhaps not surprisingly, all re-evaluations resulted in PCL-R total scores ≥ 25, suggesting at least a gross level of consistency in terms of the second assessment resulting in a "high" total score. Overall, mean scores across the two evaluations tended to be similar, although PCL-R factor and total scores were slightly higher for the second evaluation. Although none of these increases was large enough to reach statistical significance, Cohen's *d* effect sizes were at the high end of the small range for Factor 1 and total scores. Thus, despite some concerns that the way in which cases were selected for re-assessment (above-average scores relative to the group mean) might result in regression toward the mean (lower scores) when re-assessed, this was not the case. The fact that examiners were aware that a re-assessment was triggered by an elevated score on the first assessment may help to explain this finding.

Table 1. Comparisons between PCL-R factor and total scores from first and second evaluations

PCL-R score	Time 1		Time 2		*t*(19)	*d*
	M	SD	*M*	SD		
Factor 1	11.90	2.90	12.75	1.94	1.18	.34
Factor 2	13.56	2.45	13.78	2.63	0.43	.09
Total	29.29	3.04	30.66	3.53	1.76	.42

$p > .05$ for all *t*-values.

[7] It is worth clarifying that none of the PCL-R scores in this study overlaps with results reported in any other studies of PCL-R inter-rater reliability among Texas sex offenders (Boccaccini et al., 2008; Murrie et al., 2008, 2009), although a few of the individuals included in the Edens et al. (2000) data set ultimately were pursued for civil commitment by the state.

[8] For the 38 SOTP offenders with scores <25 who were not reassessed, mean PCL-R scores were total, 17.03, SD = 5.44, Factor 1, 8.31, SD = 4.05, and Factor 2, 9.36, SD = 4.46.

DOI: 10.1002/bsl

Rater Agreement

We used a generalizability theory approach to examine rater agreement for the PCL-R total and factor scores. We used a two-way random effects model for each analysis, with offenders and evaluation order (1 versus 2) as random factors. Each analysis provided variance component estimates for offenders, evaluation order, and the interaction between offenders and evaluation order. The proportion of total variance (sum of the three variance components) attributable to offenders is the absolute value intraclass correlation coefficient for a single evaluator ($ICC_{A,1}$) for the PCL-R score. This proportion is an estimate of variance in the scores that is attributable to offenders' true standing on the PCL-R. The proportion of variance attributable to evaluation order provides information about the extent to which observed score values were a product of systematic changes over time. This proportion should be near zero if score values were unrelated to evaluation order. However, if all PCL-R scores decreased over time (or all increased over time), this proportion of variance would be greater than zero and would indicate the proportion of variance attributable to evaluation order. The interaction term captures the variance in scores that was not explained by offenders or evaluation order, and is an indicator of the influence that unaccounted for sources of error have on observed scores, including random measurement error.

Table 2 provides a summary of the generalizability theory analyses. The $ICC_{A,1}$ values for the total and factor scores were well below values reported in the PCL-R manual. The ICC for the total score (.420) was similar to ICCs reported in other studies examining PCL-R rater agreement in Texas sex offender evaluations (Boccaccini et al., 2008; Murrie et al., 2009). However, the ICC for Factor 1 (.157) was notably lower than the ICC for Factor 2 (.557). Evaluation order did not account for a noteworthy proportion of variance for any of the PCL-R scores (0.0–5.5%), which is consistent with the *t*-test findings in Table 1. Thus, most of the variance that was not attributable to offenders' standing on the PCL-R was a product of unmeasured sources of error.

Range Restriction and Rater Agreement

It is well known that range restriction attenuates the size of correlations. Thus, it is likely that the rater agreement coefficients reported above are underestimates of what agreement would have been if all 58 SOTP offenders had been re-assessed with the PCL-R rather than simply those who scored ≥25 on their first assessment. To

Table 2. Rater agreement for PCL-R factor and total scores: generalizability theory

PCL-R score	Variation attributable to:		
	Offenders ($ICC_{A,1}$)	Time 1 vs. 2	Other
T			
Factor 1	.157	.017	.826
Factor 2	.557	.000	.462
Total	.420	.055	.525

$ICC_{A,1}$ = Absolute agreement intraclass correlation coefficient for a single score. $N = 20$.

estimate[9] the impact of range restriction on rater agreement in this study, we considered how range restriction impacted the Pearson correlations between raters using a range restriction correction formula provided by Cohen, Cohen, West, and Aiken (2003, p. 58).

Before describing the results of these analyses, it is necessary to point out some differences between an ICC for absolute agreement and a Pearson correlation. Although Pearson correlations can be used to measure rater agreement, they provide information about consistency agreement—not absolute value agreement. In other words, Pearson correlations consider whether the scores that evaluators give tend to rank offenders in the same order, but do not consider differences in the actual value of the score as a form of rater error. The Pearson correlation between scores from two raters is conceptually similar to a single evaluator ICC for consistency agreement, and equations for these two coefficients produce similar results in most situations (see Shrout, 1995). However, consistency agreement coefficients tend to be larger than absolute value coefficients because by definition they consider fewer sources of variance to be error (i.e., do not consider the size of the difference between two scores to be error).

The range restriction correction formula estimates a corrected correlation by considering the difference between the standard deviation of scores in the population and the standard deviation of scores in the range restricted sample. We used standard deviation values from PCL-R scores from the entire sample of 58 SOTP offenders as estimates of the population standard deviation. In other words, we used the correction formula to estimate the size of the correlation between raters if all 58 offenders had been reevaluated with the PCL-R. For each PCL-R score, the standard deviation was much larger in the entire sample (Factor 1 = 4.01, Factor 2 = 4.38, total = 7.54) than it was in the range restricted sample of 20 cases (Factor 1 = 2.90, Factor 2 = 2.45, total = 3.04). These notable differences in standard deviation values suggest that range restriction may indeed have had a substantial impact on rater agreement coefficients.

Table 3 summarizes the results of the range restriction analyses. The first column of Table 3 provides single evaluation ICC values for consistency agreement. These values are similar to the ICCs for absolute agreement in Table 2 because there was little variance associated with evaluation order (thus, most error was considered to be attributable to a combination of random measurement error and factors not measured

Table 3. Impact of range restriction on rater agreement

PCL-R score	Offenders ($ICC_{C,1}$)	Pearson *r* uncorrected	Pearson *r* corrected
Factor 1	.160	.173	.236
Factor 2	.547	.548	.761
Total	.445	.450	.781

$ICC_{C,1}$ = Consistency agreement intraclass correlation coefficient for a single score. N = 20.

[9] We want to emphasize the term "estimate" here because it assumes (most likely, erroneously) that the level of reliability seen over this particular range of scores necessarily generalizes to other score ranges on the test. Item response theory analyses would suggest that this is not the case (see Cooke & Michie, in press). Nevertheless, we believe these results are useful for illustrative purposes.

DOI: 10.1002/bsl

in the two-way ICC model). If there had been a systematic increase or decrease in scores over time, the values of the absolute and consistency agreement coefficients would have been more discrepant. The second column lists Pearson *r* values, which were, as expected, similar to the $ICC_{C,1}$ values. The final column lists Pearson correlations that were corrected for range restriction.

As expected, the *r* values that were corrected for range restriction were larger than the uncorrected *r* values. For Factor 2 and the total score, the corrected *r* values are closer in size to the values one would expect for rater agreement in applied settings (Factor 2 = .761, total score = .781). However, the corrected *r* value for Factor 1 ($r = .236$) is still well below the expected level of agreement for reliable measures.

Standard Error of Measurement

To help translate the rater agreement findings to the metric of the PCL-R, we calculated three difference scores for each offender: Factor 1, Factor 2, total score. Because evaluation order accounted for only a small amount of variance in scores, we used absolute value difference scores; that is, we subtracted the smaller score from the larger score, irrespective of which score was given first. We then calculated the percentage of difference scores that fell within and outside of what would be expected based on the standard error of measurement (SEM). About 68% of difference scores should be within one SEM unit, and 95% should be within two SEM units.

The SEM for the total score reported in the PCL-R manual is 2.90 (Hare, 2003). Although the PCL-R manual does not report SEM for the factor scores, we calculated SEM values using ICC_1 and SD values for male offenders reported in the PCL-R manual. For male offenders, the SEM for both Factor 1 and Factor 2 was about 1.9 points (1.90 for Factor 1, 1.87 for Factor 2). To put the Factor 1 value in a broader context, one would expect an average initial score (i.e., ~8 points) on a second assessment to fall somewhere between ~4 (17th percentile) and ~12 (83rd percentile) approximately 95% of the time.

Difference scores ranged from 0.00 to 8.00 for the PCL-R total score, 0.00 to 7.00 for Factor 1, and 0.00 to 6.00 for Factor 2. For both Factor 2 and the total score, 90% of the difference scores were within two SEM units (6 points) and only two (10%) difference scores were greater than 2 SEM units. However, for Factor 1, 25% of the difference scores were more than two SEM units in size, with 75% within two SEM units.

DISCUSSION

The results of this small study add to a growing body of research on the field reliability of PCL-R scores, which seems to suggest that ratings in "real world" settings may be less stable than would be presumed based on a perusal of published research studies. It is worth highlighting at the outset that the issue of field reliability is a concern in relation to numerous other types of test and rating scale that may be used in forensic settings (e.g., intelligence tests, projective instruments, DSM diagnoses). Moreover, there is growing interest more generally in the distinction between "efficacy" versus "effectiveness" in the applied assessment field (see, e.g., Mash & Hunsley, 2005), similar to the same types of distinction discussed in the "evidence-based treatment" literature.

DOI: 10.1002/bsl

That said, arguably the primary contribution of this particular assessment "effectiveness" (rather than "efficacy") study is the parsing of PCL-R scores into their two most widely studied component parts, the affective/interpersonal (or "personality") scores and the lifestyle/behavioral scores. The very poor performance of Factor 1, even after adjusting for the possible effects of range restriction and even after taking into account the already large SEM (computed based on information provided in the manual), is of considerable concern in applied settings. It suggests that the features that arguably may have the most impact on "consumers" of PCL-R results (e.g., jurors, judges) are also likely to be the most influenced by some type of measurement error.

Were the Factor 1 results in this study a completely isolated finding, it would be easier to dismiss the findings as a result of—ironically enough—random error. Such is not the case, however, in that other evidence bearing on this issue also raises concerns regarding instability across raters. First, noted earlier, the ICC values reported in the PCL-R manual for men have been consistently lower for Factor 1 than for the total score and for Factor 2—even though such findings appear to be based on single interviews that do not mimic those real-world contexts in which more than one PCL-R score is likely to be obtained. Of note, we were able to locate only a few studies that have examined inter-rater reliability of the factor scores that explicitly reported that they relied on independently conducted interviews. These studies are important in that they examine stability in a manner that is much more consistent with real-world applications of the PCL-R, similar to the present study. In an early report, Alterman, Cacciola, and Rutherford (1993) examined the one-month test–retest reliability of the PCL-R among 88 male methadone patients in which the amount of information available to the raters was systematically varied. Depending on the amount of information available, test–retest correlations ranged from .71 to .76 for Factor 1, whereas they ranged from .79 to .80 for Factor 2 and from .85 to .89 for the total score. Subsequently, Rutherford et al. (1999) reported ICC values for PCL-R scores in a much larger sample of male methadone patients ($n = 200$) participating in their longitudinal research. At a two-year follow-up, again involving independent evaluations and new interviews, Factor 1 scores were significantly less reliable (ICC = .43) than were Factor 2 scores (ICC = .60) or total scores (ICC = .60).[10] More recently, Tyrer et al. (2005) examined PCL-R reliability in a sample of 15 British offenders housed in a high security hospital over an average of an 11-month follow-up. Inter-rater agreement (ICC) in this small male sample was .59 for the total score but only .49 for Factor 1 (and .44 for Factor 2). Although obviously confounded with the length of the follow-up period, the most important issue in relation to the current study is the poorer performance of Factor 1 in these studies relative to the total score. In this context, the results obtained in the present study do not appear to be especially out of the ordinary.

Moreover, the results of our analyses of difference scores should also be considered in the context of recent Monte Carlo work examining the stability of PCL-R scores based on SEMs reported in the instrument's manual. Cooke and Michie (in press) have argued persuasively that, because SEMs are based on the erroneous assumption that the degree of reliability is static across all score levels on a rating scale, it is more appropriate to consider the conditional SEM (CSEM). The CSEM accounts for the fact that reliability is not a constant at different levels of a test but in fact becomes worse as scores

[10] Values for the 25 women in this study were .65 (total), .63 (Factor 1), and .50 (Factor 2).

DOI: 10.1002/bsl

become more extreme. This is also true for the factor scores, which again may help to explain the large Factor 1difference scores obtained in our study that were beyond what one would expect based on the SEM that can be computed from the manual data. One would expect rater differences at higher score values to be somewhat more pronounced—although not as pronounced as what was observed here.

The amount of unexplained variance in Factor 1 scores in this and other studies raises the interesting question of what exactly *does* explain this variability. Murrie and colleagues have posed adversarial allegiance as one possible source of systematic error, but that would seem less relevant in a study such as ours in which all examiners were state employees. Following up on their earlier findings suggesting the possibility of adversarial allegiance, Boccaccini et al. (2008) recently reported analyses of a much larger sample of PCL-R scores, all of which were obtained from state-retained experts conducting SVP evaluations. They found that about 30% of the variance in PCL-R total scores was attributable to the individual rater completing the PCL-R, which was much higher than would be expected based on existing estimates of inter-rater reliability. Indeed, if the ICC for the PCL-R was .80, no more than 20% of the variance could be attributable to any other source. In addition, the mean PCL-R scores reported by commonly employed examiners varied to a considerable degree. For example, the most extreme mean total score difference across two examiners who had each conducted at least 10 evaluations for the state was over 14 points ($M = 31.75$ versus $M = 17.50$). As such, one partial explanation for the poor performance of Factor 1 in our study is that individual raters' own subjective thresholds for judging the level of Factor 1 traits may vary widely. Some examiners may require, in essence, a "lower bar" to assign a score of 2 on certain key items, compared with others who may use a generally higher bar before assigning a value of 2.

Alternatively, some evaluators may have their own idiosyncratic beliefs about the types of behavior that do and do not warrant certain scores on these items. This may be particularly true in relation to the assessment of sex offenders (especially those targeting children), in that it may be difficult to infer reliably the presence of certain affective and interpersonal characteristics because they may vary considerably depending on the context in which they are considered. For example, a pedophile might seem exceedingly "superficially charming" in one context (e.g., interactions with child victims) and yet otherwise seem to lack this particular characteristic. Similarly, "acceptance of responsibility" for sexual conduct with children may be an especially problematic determination to make, particularly among individuals who have been indoctrinated into what they believe they are *supposed to say* to prison staff. (For a similar discussion of some of the difficulties of rating personality traits on the PCL-R in relation to substance abusers, see Rutherford et al., 1999.)

Moreover, it is certainly possible that different evaluators actually *evoke* different levels of Factor 1 traits from offenders due to their own interviewing styles. For example, individual differences in the interpersonal styles of interviewers may elicit different reactions, such as a "high dominance/low nurturance" examiner suppressing the amount of glibness and superficial charm that might be evident in an interview with an examiner with more of a "low dominance/high nurturance" interviewing style. Although interview data should not be the sole information upon which the scoring of these items is based, it clearly informs these ratings. According to Hare (Hare, 2003, pp. 57–58), Factor 1 ratings based only on file review are on average about two points lower than those including interview data.

Finally, perhaps it should not be forgotten that interview-based psychiatric diagnoses in general have a less-than-stellar history in terms of reliability and that one of the primary reasons for moving away from the more inferential constructs in DSM-II to the more behaviorally based criteria evident in DSM-III (and later editions) was to minimize exactly the types of inconsistency that are evident in our results for Factor 1 (see, generally, Lilienfeld, 1998). As noted earlier, the PCL-R manual eschews the use of very structured interview protocols (see Hare, 2003, which states (in boldface) "Users should try to avoid interviews that are highly structured" p. 18), in large part to discourage highly reliable yet potentially less valid and useful information from being obtained. Paradoxically, such an approach may lead to highly reliable ratings when *based on the same interview*, yet may result in much less stable ratings on Factor 1 across *separate interviews*, such as in the case of our results (and those of Rutherford et al., 1999, and Tyrer et al., 2005).

Noted earlier, Rufino et al. (2009) recently reported that many PCL-R studies involving interviews do not clearly indicate how these are conducted (joint or separate) in relation to the collection of inter-rater reliability data. Given the added labor of separate interviews, it seems reasonable to presume that at least in most research studies only one was conducted. Given the potential importance of this design issue, however, it would certainly behoove researchers working in this area to be much clearer about their methods when describing the results of their reliability studies and analyses. Of note, it would also be exceedingly helpful if researchers would clarify the specific nature of the statistics being reported. Some studies simply report a vague "correlation" or "inter-rater reliability" value that could either be a Pearson *r* or an ICC. Even when clear that an ICC was used, many studies are not explicit as to whether the value is for consistency or absolute agreement, or whether the results are for a single or multiple raters—all of which are factors that can significantly impact the interpretation of the reported correlational value. We would encourage researchers to report ICC results using the coefficient identifiers explained by McGraw and Wong (1996).

The limitations of this particular study are fairly self-evident and will only be briefly highlighted here. Obviously, it would have been preferable had reliability data been available for all 58 inmates who were initially assessed rather than a subset of individuals who initially obtained high scores. This concern is attenuated somewhat by our range restriction analyses but future research would benefit from analyses of the full range of "typical" PCL-R scores in offender samples. Similarly, given interest in recent three- and four-factor models of the PCL-R, it would be beneficial to examine the reliability of these factors, as well as (given a large enough sample) item level reliability statistics such as those reported by Rutherford et al. (1999) and in the PCL-R manual (Hare, 2003). Additionally, more information concerning the individual raters would allow for more sophisticated analyses, such as further parsing of the unexplained variance into its component sources (e.g., individual rater effects). In most studies, researchers can use a generalizability framework to estimate the proportion of variance in test scores attributable to individual raters, evaluation order, or other reliability study design features (Brennan, 2001). Even better would be information concerning the training of examiners and whether this has a substantive impact on the reliability of the scores they produce. For example, it is conceivable that training may be more or less useful in improving reliability on certain items or factors. Finally, it obviously would be beneficial to expand field reliability studies beyond sex offender samples, given that the PCL-R is widely used in other types of criminal cases (DeMatteo & Edens, 2006).

DOI: 10.1002/bsl

These limitations notwithstanding, the present results add to a growing body of research raising concerns about the reliability of PCL-R scores across examiners in field settings and in particular highlight the potentially problematic nature of those items historically associated with the "personality" components of this rating scale (i.e., Factor 1). Given the increasing use of the PCL-R and related scales "outside the lab," we hope that this research stimulates other researchers to examine this important issue in other applied settings.

REFERENCES

Alterman, A., Cacciola, J., & Rutherford, M. (1993). Reliability of the Revised Psychopathy Checklist in substance abuse patients. *Psychological Assessment*, *5*, 442–448.

Archer, R. P., Buffington-Vollum, J. K., Stredny, R. V., & Handel, R. W. (2006). A survey of psychological test use patterns among forensic psychologists. *Journal of Personality Assessment*, *87*, 84–94.

Boccaccini, M. T., Turner, D. T., & Murrie, D. C. (2008). Do some evaluators report consistently higher or lower scores on the PCL-R?: Findings from a statewide sample of sexually violent predator evaluations. *Psychology, Public Policy, and Law*, *14*, 262–283.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Cohen, J., Cohen, P., West, S., & Aiken, L. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lowrence Erlbaum Associates.

Cooke, D. J., & Michie, C. (in press). Limitations of diagnostic precision and predictive utility in the individual case: A challenge for forensic practice. *Law and Human Behavior.*

Cooke, D. J., Michie, C., & Skeem, J. L. (2007). Understanding the structure of the Psychopathy Checklist—Revised. *British Journal of Psychiatry*, *190*(suppl 49), s39–s50.

Costanzo, M., & Peterson, J. (1994). Attorney persuasion in the capital penalty phase: A content analysis of closing arguments. *Journal of Social Issues*, *50*, 125–147.

DeMatteo, D., & Edens, J. F. (2006). The role and relevance of the Psychopathy Checklist—Revised in court: A case law survey of U.S. courts (1991–2004). *Psychology, Public Policy, and Law*, *12*, 214–241.

Edens, J. F. (2006). Unresolved controversies concerning psychopathy: Implications for clinical and forensic decision-making. *Professional Psychology: Research and Practice*, *37*, 59–65.

Edens, J. F., Campbell, J. S., & Weir, J. M. (2007). Youth psychopathy and criminal recidivism: A meta-analysis of the Psychopathy Checklist measures. *Law and Human Behavior*, *31*, 53–75.

Edens, J. F., Colwell, L. H., Desforges, D. M., & Fernandez, K. (2005). The impact of mental health evidence on support for capital punishment: Are defendants labeled psychopathic considered more deserving of death? *Behavioral Sciences and the Law*, *23*, 603–625.

Edens, J. F., Hart, S. D., Johnson, D. W., Johnson, J., & Olver, M. E. (2000). Use of the Personality Assessment Inventory to assess psychopathy in offender populations. *Psychological Assessment*, *12*, 132–139.

Edens, J. F., Marcus, D. K., Lilienfeld, S. O., & Poythress, N. G. (2006). Psychopathic, not psychopath: Taxometric evidence for the dimensional structure of psychopathy. *Journal of Abnormal Psychology*, *115*, 131–144.

Edens, J. F., & Petrila, J. (2006). Legal and ethical issues in the assessment and treatment of psychopathy. In C. Patrick (Ed.), *Handbook of psychopathy* (pp. 573–588). New York: Guilford.

Edens, J. F., Petrila, J., & Buffington-Vollum, J. K. (2001). Psychopathy and the death penalty: Can the Psychopathy Checklist—Revised identify offenders who represent "a continuing threat to society?'. *Journal of Psychiatry and Law*, *29*, 433–481.

Edens, J. F., & Vincent, G. M. (2008). Juvenile psychopathy: A clinical construct in need of restraint? *Journal of Forensic Psychology Practice*, *8*, 186–197.

Forth, A. E., Kosson, D. S., & Hare, R. D. (2003). *The Psychopathy Checklist: Youth Version*. Toronto, Ontario: Multi-Health Systems.

Gendreau, P., Goggin, C., & Smith, P. (2002). Is the PCL-R really the "unparalleled" measure of offender risk? A lesson in knowledge cumulation. *Criminal Justice and Behavior*, *29*, 397–426.

Hare, R. D. (1991). *The Hare Psychopathy Checklist—Revised manual*. North Tonawanda, NY: Multi-Health Systems.

Hare, R. D. (2003). *The Hare Psychopathy Checklist—Revised manual* (2nd ed.). North Tonawanda, NY: Multi-Health Systems.

Hare, R. D. (2006). Psychopathy: A clinical and forensic overview. *Psychiatric Clinics of North America*, *29*, 709–724.

Hare, R. D., & Neumann, C. S. (2005). The structure of psychopathy. *Current Psychiatry Reports*, *7*, 57–64.

DOI: 10.1002/bsl

Harris, G. T., Rice, M. E., & Quinsey, V. L. (1994). Psychopathy as a taxon: Evidence that psychopaths are a discrete class. *Journal of Consulting and Clinical Psychology*, *62*, 387–397.

Hemphill, J. F., & Hare, R. D. (2004). Some misconceptions about the Hare PCL-R and risk assessment: A reply to Gendreau, Goggin, and Smith. *Criminal Justice and Behavior*, *31*, 203–243.

Lally, S. J. (2003). What tests are acceptable for use in forensic evaluations? A survey of experts. *Professional Psychology: Research and Practice*, *5*, 491–498.

Levenson, J. S. (2004). Reliability of sexually violent predator civil commitment criteria in Florida. *Law and Human Behavior*, *28*, 357–368.

Lilienfeld, S. (1998). Methodological advances and developments in the assessment of psychopathy. *Behaviour Research and Therapy*, *36*, 99–125.

Lloyd, C. D., Clark, H. J., & Forth, A. E. (in press). Psychopathy, expert testimony and indeterminate sentences: Exploring the relationship between Psychopathy Checklist—Revised testimony and trial outcome in Canada. *Legal and Criminological Psychology*.

Mash, E. J., & Hunsley, J. (2005). Evidence-based assessment of child and adolescent disorders: Issues and challenges. *Journal of Clinical Child and Adolescent Psychology*, *34*, 362–379.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30–46.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749.

Murrie, D. C., Boccaccini, M., Johnson, J., & Janke, C. (2008). Does interrater (dis)agreement on Psychopathy Checklist scores in sexually violent predator trials suggest partisan allegiance in forensic evaluation. *Law and Human Behavior*, *32*, 352–362.

Murrie, D. C., Boccaccini, M., Turner, D., Meeks, M., Woods, C., & Tussey, C. (2009). Rater (dis)agreement on risk assessment measures in sexually violent predator proceedings: Evidence of adversarial allegiance in forensic evaluation? *Psychology, Public Policy, and Law*, *15*, 19–53.

Nicholls, T. L., & Petrila, J. (2005). Gender and psychopathy: An overview of important issues and introduction to the special issue. *Behavioral Sciences and the Law*, *23*, 729–741.

Rufino, K., Heinonen, L., Boccaccini, M. T., Murrie, D. C., & Edens, J. F. ((2009). March). *What do PCL rater-agreement coefficients tell us about forensic practice?* Paper presented at the annual meeting of the American Psychology-Law Society, San Antonio, TX.

Rutherford, M., Cacciola, J. S., Alterman, A. I., McKay, J. R., & Cook, T. G. (1999). The 2-year test–retest reliability of the Psychopathy Checklist—Revised in methadone patients. *Assessment*, *6*, 285–291.

Shrout, P. E. (1995). Measuring the degree of consensus in personality judgements. In P. E. Shrout and S. T. Fiske (Eds.), *Personality research, methods, and theory: A Festschrift honoring Donald W. Fiske* (pp. 79–92). Hillsdale, NJ: Lowrence Erlbaum Associates.

Skeem, J., & Cooke, D. (in press). Is antisocial behavior essential to psychopathy? Conceptual directions for resolving the debate. *Psychological Assessment*.

Skeem, J. L., Edens, J. F., Camp, J., & Colwell, L. H. (2004). Are there racial differences in levels of psychopathy? A meta-analysis. *Law and Human Behavior*, *28*, 505–527.

Skeem, J. L., Monahan, J., & Mulvey, E. P. (2002). Psychopathy, treatment involvement, and subsequent violence among civil psychiatric patients. *Law and Human Behavior*, *26*, 577–603.

Sullivan, E., & Kosson, D. (2006). Ethnic and cultural variations in psychopathy. In C. Patrick (Ed.), *Handbook of psychopathy* (pp. 437–458). New York: Guilford.

Sundby, S. E. (1998). The capital jury and absolution: The intersection of trial strategy, remorse, and the death penalty. *Cornell Law Review*, *83*, 1557–1598.

Texas Health & Safety Code § 841.000–841.150 (2000).

Tyrer, P., Cooper, S., Seivewright, H., Duggan, C., Rao, B., & Hogue, T. (2005). Temporal reliability of psychological assessments for patients in a special hospital with severe personality disorder: A preliminary note. *Criminal Behaviour and Mental Health*, *15*, 87–92.

Verona, E., & Vitale, J. (2006). Psychopathy in women: Assessment, manifestations, and etiology. In C. Patrick (Ed.), *Handbook of psychopathy* (pp. 415–436). New York: Guilford.

Wood, J., Nezworski, M., & Stejskal, W. (1996). The Comprehensive System for the Rorschach: A critical examination. *Psychological Science*, *7*, 3–10.

DOI: 10.1002/bsl